AD-A100 146 OLD DOMINION UNIV NORFOLK VA DEPT OF MATHEMATICAL SC--ETC F/G 12/1
MINIMAX RIDGE REGRESSION.(U)
MAY 80 L C PEELE, T P RYAN F49620-79-C-0125
UNCLASSIFIED TR-1 AFOSR-TR-80-1092 NL

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER AFOSR-TR- 80-1092	2. GOVT ACCESSION NO. AD-A100146
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle) MINIMAX RIDGE REGRESSION	5. TYPE OF REPORT & PERIOD COVERED FINAL
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) L. PEELE T.P. RYAN	8. CONTRACT OR GRANT NUMBER(s) F49620-79-C-0125
9. PERFORMING ORGANIZATION NAME AND ADDRESS Dept of Mathematics Sciences Old Dominion University Norfolk, VA	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F 2304/A5
11. CONTROLLING OFFICE NAME AND ADDRESS AFOSR/NM Bldg 410 Bolling AFB, DC 20332	12. REPORT DATE May 80 13. NUMBER OF PAGES 22
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) LEVEL II	15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for Public Release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This work examined minimax linear estimation in multiple linear regression. The application of minimax estimation to regression led to the development of ridge regression estimators with stochastic ridge parameters. These estimators were seen to be invariant under linear transformation; a property which has not been established for other ridge estimators. These minimax-motivated estimators were examined in several simulation studies. In particlar, flaws in other simulation studies of ridge estimators were depicted. (Continued)

Unclassified 411864

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

DISCLAIMER NOTICE

THIS DOCUMENT IS BEST QUALITY PRACTICABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

20. Cont.

Consequently, an improved simulation procedure was used. It was observed from these studies that, contrary to published statements, a ridge estimator can be considerably superior to the ordinary least squares estimator, especially when high pairwise correlations exist among the regression variables. Robustness considerations were used to suggest a requirement that a "good" generalized ridge regression estimator should satisfy.

AFOSR-TR- 80-1092

[redacted]

MINIMAX RIDGE REGRESSION

(Final Report)

ABSTRACT

This work examined minimax linear estimation in multiple linear regression. The application of minimax estimation to regression led to the development of ridge regression estimators with stochastic ridge parameters. These estimators were seen to be invariant under linear transformation; a property which has not been established for other ridge estimators. These minimax-motivated estimators were examined in several simulation studies. In particular, flaws in other simulation studies of ridge estimators were depicted. Consequently, an improved simulation procedure was used. It was observed from these studies that, contrary to published statements, a ridge estimator can be considerably superior to the ordinary least squares estimator, especially when high pairwise correlations exist among the regression variables. Robustness considerations were used to suggest a requirement that a "good" generalized ridge regression estimator should satisfy.

Activities partially supported by AFOSR contract:

1. "The Merits of some Ridge Regression Estimators," research by Lawrence Peele and Thomas P. Ryan presented by Dr. Ryan at the 1979 joint ASA, IMS, Biometric Society meetings.

2. "Minimax Linear Regression Estimators with Application to Ridge Regression," by Lawrence Peele and Thomas P. Ryan, to appear in Technometrics.

3. "Comment" (on invited ridge regression paper) by Lawrence Peele and Thomas P. Ryan, J. Amer. Statist. Assoc. 75 96-97.

4. "Most-robust Ridge Regression Estimators," by Lawrence Peele, submitted for publication.

MINIMAX LINEAR REGRESSION ESTIMATORS WITH APPLICATION TO RIDGE REGRESSION[1]

by

Lawrence Peele and Thomas P. Ryan

AFOSR Technical Report No.1

May, 1980

Old Dominion University
Department of Mathematical Sciences
Norfolk, Virginia

[1] To appear in Technometrics. Research partially supported by AFSC under Contract F49620-79-C-0125.

ABSTRACT

This paper considers minimax linear estimation of the parameters in a multiple linear regression model. Recent results are summarized, and some new results, including a transformation invariance property of minimax estimation, are given. These minimax estimators of the parameter vector can also be classified as ridge regression estimators with nonstochastic ridge parameters. Some ridge regression estimators with stochastic ridge parameters can be motivated by minimax estimation considerations. These minimax-motivated estimators are examined in several simulation studies and some observations are made based on these simulations and minimax theory.

1. INTRODUCTION

The usual multiple linear regression model is

$$Y = X\beta + \xi$$

where ξ is a vector of uncorrelated random variables with mean zero and variance σ^2, and X is a full rank $n \times q$ matrix with $q < n$. The usual procedure for estimating β is the least squares method. It is well known that this method is equivalent to minimum variance unbiased linear (MVUL) estimation. Similarly, the usual method for estimating a given linear combination of the coefficients is MVUL. In recent years, many people have attempted to reduce the mean squared error (MSE) by allowing some bias in their estimators. One such biased estimation procedure is ridge regression, which was first studied by Hoerl and Kennard [7]. Ridge regression estimators have the form

$$\hat{\beta}^*(k) = [X'X + kI]^{-1}X'Y$$

where the ridge parameter k is either nonstochastic (based on prior information) or stochastic. It follows from a result in [7] that for nonstochastic k satisfying

$$k < \sigma^2/\beta'\beta \qquad (1)$$

the ridge regression estimator $\hat{\beta}^*(k)$ outperforms the least squares estimator in terms of MSE. The minimax linear estimators discussed in the present paper are based on constraints related to (1). These estimators are shown (originally in [9]) to be ridge regression estimators, and both known and new properties of minimax linear regression estimators are given. Also, some simulation results are presented with an accompanying discussion.

2. MINIMAX LINEAR ESTIMATION

In general, minimax procedures "guard against the worst." It is observed in Section 3 that such caution is appropriate in highly ill-conditioned regression problems.

The following lemma is the basis for minimax linear estimation of β. The Cauchy-Schwarz inequality (See, for example, page 215 of [12].) when applied to R^q results in the following lemma.

Lemma 1. Let $a \in R^q$. The maximum of $(a'r)^2$ among $r \in R^q$ satisfying $r'r \leq 1$ is $a'a$.

Let $k > 0$. Let A be a positive definite (p.d.) matrix of order q, and consider the constraint

$$\beta' A\beta/\sigma^2 \leq k. \qquad (2)$$

Let $a \in R^q$ and suppose that we wish to estimate $a'\beta$. If $\tilde{\beta}$ satisfies

$$\min_{c} \max_{\beta,\sigma \text{ in } (2)} E_{\beta,\sigma}[(c'Y - a'\beta)^2]$$

$$= \max_{\beta,\sigma \text{ in } (2)} E_{\beta,\sigma}[(\tilde{c}'Y - a'\beta)^2]$$

then $\tilde{c}'Y$ is the minimax linear estimator of $a'\beta$ based on (2).

The following result is presented in [9]. A brief proof is given as the method of proof seems useful.

Theorem 1. The minimax linear estimator $\tilde{c}'Y$ of $a'\beta$ based on (2) satisfies $\tilde{c}'Y = a'\hat{\beta}_{A,k}$ where

$$\hat{\beta}_{A,k} = [X'X + (1/k)A]^{-1}X'Y$$

Proof. Since A is p.d. we can factor A as $A = S'S$. Condition (2) becomes $\gamma'\gamma \le \sigma^2 k$ where $S\beta = \gamma$. Hence, $\tilde{c}$ must minimize

$$\max_{\gamma'\gamma \le \sigma^2 k} E_{\gamma,\sigma}[(c'Y - a'S^{-1}\gamma)^2]$$

$$= \sigma^2 c'c + \max_{\gamma'\gamma \le \sigma^2 k}[c'XS^{-1}\gamma - a'S^{-1}\gamma]^2$$

$$= \sigma^2 c'c + \sigma^2 k \max_{\gamma'\gamma \le 1}[(c'XS^{-1} - a'S^{-1})\gamma]^2$$

$$= \sigma^2 c'c + \sigma^2 k(c'X - a')A^{-1}(X'c - a) \qquad (3)$$

by Lemma 1. By finding the zero of the vector derivative of (3) with respect to c, we can see that the c which minimizes (3) is given by

$$\begin{aligned} c &= [(1/k)I_n + XA^{-1}X']^{-1}XA^{-1}\alpha \\ &= X[(1/k)I_p + A^{-1}X'X]^{-1}A^{-1}\alpha \\ &= X[X'X + (1/k)A]^{-1}\alpha \end{aligned}$$

(since $X[(1/k)I_p + A^{-1}X'X] = [(1/k)I_n + XA^{-1}X']X$)
where the subscript of I denotes the order of the identity matrix. The theorem follows immediately, since the Hessian of (3) is always positive definite.

It follows from Theorem 1, by letting α be the Kronecker delta δ_i, that minimax estimation of individual components β_i of β based on (2) also produces the estimator $\hat{\beta}_{A,k}$ of β.

One can easily verify that the minimax linear estimator $\hat{\beta}_{A,k}$ based on (2) is equivalent to the minimax linear estimator based on

$$\beta'A\beta/\sigma^2 = k. \qquad (4)$$

Hence, minimax theory suggests that if $\hat{k}$ is an estimator of $\beta'A\beta/\sigma^2$, then one might use

$$\hat{\beta}_{A,\hat{k}} = (X'X + (1/\hat{k})A)^{-1}X'Y$$

to estimate k. One could also argue that overestimation of k by $\hat{k}$ seems preferable to underestimation since, in (2), k is an upper bound; however, robustness results discussed in [11] and simulation results given in the present paper indicate that overestimation may result in unnecessarily and undesirably cautious estimators.

Example 1. Consider the case when $A = I_q$. The minimax linear estimator $\hat{\beta}_{I,k}$ is identical to the ridge regression estimator $\beta^*(\lambda)$ where $\lambda = k^{-1}$. Minimax theory supports the estimator $\hat{\beta}_{I,\hat{k}_1}$ where $\hat{k}_1 = \hat{\beta}'\hat{\beta}/\hat{\sigma}^2$ and $\hat{\beta}$ and $\hat{\sigma}^2$ are the usual estimators of β and σ^2 given by $\hat{\beta} = (X'X)^{-1}X'Y$ and $\hat{\sigma}^2 = (Y-X\hat{\beta})'(Y-X\hat{\beta})/(n-q)$.

Example 2. Let $A = X'X$. Since A is p.d., we can write $A = C'C$ where C is a nonsingular $q \times q$ matrix. Hence, the model $Y = X\beta + \varepsilon$ can be written as $Y = \tilde{X}\theta + \varepsilon$ where $\tilde{X} = XC^{-1}$ and $\theta = C\beta$. Since $\tilde{X}'\tilde{X} = I$, it seems reasonable to place equal weight on the size of different components of θ; that is, the constraint $\theta'\theta \le \sigma^2 k$ seems very reasonable. This constraint is equivalent to the constraint $\beta'X'X\beta/\sigma^2 \le k$ which leads to $\hat{\beta}_{X'X,k} = [X'X + (1/k)X'X]^{-1}X'Y = (1 + k^{-1})^{-1}\hat{\beta}$. This estimator is a shrunken estimator. Estimation of k by $\hat{k} = \hat{\beta}'X'X\hat{\beta}/\hat{\sigma}^2$ leads to the estimator $(1 + \hat{k}^{-1})^{-1}\hat{\beta}$ which is a shrunken estimator with a stochastic shrinkage coefficient.

In [illegible], Kuks and Olman also consider minimax linear regression estimators based on assumptions of the form

$$(\beta-\beta^0)'A(\beta-\beta^0)/\sigma^2 \le k$$

where A, σ^2, and k are known. Some of the results from [9] and some additional related results are given in [2]. See Chapter 10 of [1] for a detailed, translated discussion of the results of [9].

Kuks and Olman show in [9] that for any $a \in R^q$,

$$E_{\beta,\sigma}[(a'\hat{\beta}_{A,k} - a'\beta)^2] \le E_{\beta,\sigma}[(a'\hat{\beta} - a'\beta)^2],$$

for all β,σ satisfying $\beta'A\beta/\sigma^2 \le k$. A generalization in [11] of a result from [15] states that for any nonnegative definite (n.d.) $q \times q$ matrix W,

$$E_{\beta,\sigma}[(\hat{\beta}_{A,k} - \beta)'W(\hat{\beta}_{A,k} - \beta)] \le E_{\beta,\sigma}[(\hat{\beta} - \beta)'W(\hat{\beta} - \beta)],$$

for all β,σ satisfying $\beta'A\beta/\sigma^2 \le k$. Peele observes in [11] that the two above-mentioned results, which imply robustness of $\hat{\beta}_{A,k}$ relative to assumption (2), are equivalent and notes that the A matrices in Examples 1 and 2 are optimal (in a most-robust sense) prior information matrices. Clearly the accuracy of $\hat{\beta}_{A,k}$ as an estimator of β will depend on the accuracy of the prior information (2) which involves A and k.

The following invariance property of minimax linear estimation supports the model-transformation argument used in Example 2 and also answers a criticism of (non-minimax) ridge regression techniques made in [14].

Theorem 7. Let N be any $q \times q$ nonsingular matrix. It follows that the model $Y = X\beta + \xi$ is equivalent to the model $Y = \tilde{X}\gamma + \xi$ where $\tilde{X} = XN$ and $\gamma = N^{-1}\beta$. Also, the condition $\beta'A\beta/\sigma^2 \leq k$ is equivalent to the condition $\gamma'N'AN\gamma/\sigma^2 \leq k$ since $\beta = N\gamma$. Minimax linear estimation is invariant under model transformations in that $\hat{\beta}_{A,k} = N\hat{\gamma}_{N'AN,k}$.

Proof. One can easily check that

$$\begin{aligned} N\hat{\gamma}_{N'AN,k} &= N[\tilde{X}'\tilde{X} + (1/k)N'AN]^{-1}\tilde{X}'Y \\ &= N[N'X'XN + (1/k)N'AN]^{-1}N'X'Y \\ &= [X'X + (1/k)A]^{-1}X'Y \\ &= \hat{\beta}_{A,k}. \end{aligned}$$

The following theorem appears to answer a question posed in Chapter 11 of [1] concerning the existence of a characterization of minimax linear regression estimation similar to the least squares characterization of MVUL regression estimators.

Theorem 8. Among all $b \in R^q$, $b = \hat{\beta}_{A,k}$ minimizes

$$(1/k)b'Ab + (Y - Xb)'(Y - Xb). \tag{5}$$

Proof. The zero of the vector derivative of (5) with respect to b is $b = \hat{\beta}_{A,k}$, and the Hessian of (5) is always positive definite.

An immediate consequence of Theorem 3 is the result that among b satisfying $b'Ab = (\hat{\beta}_{A,k})'A(\hat{\beta}_{A,k})$, $(Y - Xb)'(Y - Xb)$ is minimized by $b = \hat{\beta}_{A,k}$. This result has been previously observed in [7] for the case $A = I$ by Lagrange multiplier methods. See [16] for a discussion of results similar to Theorem 3.

One can easily check by orthogonal transformation methods similar to those on page 62 of [7] that if

$$g_1(k) = [\hat{\beta}^*(k)]'[\hat{\beta}^*(k)]$$

then g_1 is a decreasing function of k. Hence, it follows easily (by supposing not and then contradicting Theorem 3) that if

$$g_2(k) = [Y - X\hat{\beta}^*(k)]'[Y - X\hat{\beta}^*(k)]$$

then g_2 is an increasing function of k.

Two loss functions different from $g_2(k)$, which is minimized by the least squares estimator $\hat{\beta} = \hat{\beta}^*(0)$, are considered in Section 3 and simulation results are given for minimax-related ridge regression estimators. The simulations involve progressive ill-conditioning and some of the simulations use random, norm-one β.

3. NUMERICAL RESULTS

Within the past five years, many ridge estimators have appeared in the literature and have been compared in subsequent simulation studies. Golub, Heath, and Wahba [5] state that two dozen is probably a conservative estimate of the number of such estimators.

Many practitioners have used the ridge-trace approach to select the value of the ridge constant. This approach is highly subjective, however, and has also been criticized on other grounds (see, e.g., Smith and Campbell [14]). Consequently, it seems as though more formal methods for selecting the ridge constant need to be employed. The ridge estimator presented in Section 1,

$$\hat{\beta}^*(k_1) = (X'X + k_1 I)^{-1}X'Y$$

where k_1 is stochastic and equal to $\hat{\sigma}^2/\hat{\beta}'\hat{\beta}$, has been the focal point in our simulation studies. Some of these results are presented later in this section. We have also studied the shrinkage estimator presented in Example 2, but it did not fare as well as the ridge estimator.

Before undertaking a simulation study of regression estimators, an experimenter must select an appropriate loss function, and choose a mechanism for generating data that is representative of actual data. The loss function most frequently used is squared error, i.e.

$$(\tilde{\beta}-\beta)'(\tilde{\beta}-\beta) \qquad (6)$$

where $\tilde{\beta}$ denotes an estimator of β (see, e.g. Gibbons [4] and Hemmerle and Brantle [6]).

We have used (6) in addition to the loss function

$$(\tilde{\beta}-\beta)'X'X(\tilde{\beta}-\beta), \qquad (7)$$

which was also used by Dempster, Schatzoff, and Wermuth [3]. It can be argued (as they did) that (6) would be an appropriate loss function if the primary goal of a regression study is estimation of the parameters, whereas (7) would be appropriate if the regression was to be used for prediction (since (7) may be written as $(X\tilde{\beta}-X\beta)'(X\tilde{\beta}-X\beta)$, and $Y = X\beta + \epsilon$ is predicted by $\hat{Y} = X\tilde{\beta}$). The loss function $(Y-X\tilde{\beta})'(Y-X\tilde{\beta})$ would be inappropriate for either estimation or prediction since the loss is minimized when $\tilde{\beta}$ is the least squares estimator, as was mentioned in Section 2.

If we adopt either (6) or (7) or both, we then must decide what X and β to use as well as the method for generating values of Y. Newhouse and Oman [10] showed that, assuming X, σ^2, and k to be fixed, $MSE(\hat{\beta}^*(k))$ is maximized, for $\beta'\beta = 1$, when β is the normalized eigenvector corresponding to the smallest eigenvalue of $X'X$. Similarly, $MSE(\hat{\beta}^*(k))$ is minimized when β is the normalized eigenvector corresponding to the largest eigenvalue of $X'X$. This result has apparently led researchers to use these two choices for β in their simulation studies. Typically, X is selected in such a way as to make $X'X$ an equicorrelation matrix, and observations on the dependent variable are then generated as

$$Y = X\beta + e \tag{8}$$

where e is $N(0,\sigma^2)$. Several values of σ^2 are then used in conjunction with each of several X matrices. The loss for each estimator is then determined using either (6) or (7), or both.

This general procedure has two major shortcomings. First, we would not expect to encounter an equicorrelation matrix with actual data. Second, although the two choices for β correspond to the best and to the worst case for ridge regression estimators (in terms of mean squared error), we advise against using the smallest eigenvalue case. The reason can be discerned from Table 1. In particular, we should notice what happens to the average value of R^2 (the coefficient of multiple determination) as we move, keeping σ^2 fixed, from the well-conditioned matrix in the first part of the table, to the highly ill-conditioned matrix in the bottom part of the table. One can observe (as in Table 1) that R^2 will be approximately q/n in the smallest eigenvalue case even for moderate σ^2 if $X'X$ is highly ill-conditioned. Since this is the worst case for ridge estimators, we are thus unable to see how poorly a ridge estimator performs relative to least squares for reasonable values of R^2. Since progressive ill-conditioning tends to cause R^2 for the smallest eigenvalue case to be much different from R^2 for the largest eigenvalue case, making σ^2 smaller so as to produce higher R^2 values for the smallest eigenvalue case would tend to make R^2 almost exactly 1.0 in the largest eigenvalue case.

For these reasons, we have used a different simulation procedure. We have also generated observations on Y as in (8), but, for each trial, a β vector is generated from a uniform distribution on the collection of all norm-one q-element vectors. A normal$(0,\sigma^2)$ error vector e is then generated and Y is computed as $Y = X\beta + e$. Thisted [16] cautions against the use of random β in simulation studies which include different classes of estimators. We see nothing wrong, however, with using β vectors that are uniform on the collection of all norm-one vectors when comparing estimators within a particular class such as ridge estimators. As will be seen later, the main objectives of our numerical studies have been: (1) to compare our ridge estimator with the much-referenced Hoerl, Kennard, and Baldwin [8] estimator, and (2) to determine under what conditions our ridge constant or a multiple of our ridge constant would be appropriate.

In addition to our different procedure for generating β, we have also generated $X'X$ matrices in a different manner. Instead of the equicorrelation approach, we used the method given in Ryan [13] to generate several progressively ill-conditioned matrices.

The results of our major simulation study are shown in Table 2. The following can be discerned from inspection of the table. For a particular degree of ill-conditioning, the size of the desired ridge constant depends on the size of R^2 in that a large R^2 implies the need for a small ridge constant, and vice versa. Secondly, for fixed R^2, the size of the desired ridge constant varies with the degree of ill-conditioning in that greater ill-conditioning implies the need for a larger ridge constant.

It follows from the discussion in Section 2 concerning robustness that the use of a nonstochastic ridge constant k will result in the ridge estimator outperforming least squares in terms of MSE provided that $k < (h/2)^{-1} = 2h^{-1}$ where $h = \beta'\beta/\sigma^2$. Consequently, $2\hat{h}_1^{-1} = 2\hat{\sigma}^2/\hat{\beta}'\hat{\beta}$ (in addition to $\hat{h}_1^{-1}$ given in Example 1) should be a good (stochastic) choice for k. This choice can be motivated, in part, by the results in Table 2, and also by the fact that $E(\hat{\sigma}^2/\hat{\beta}'\hat{\beta})$ should be less than $E(\sigma^2/\beta'\beta)$ since (assuming normality) $E(\hat{\beta}'\hat{\beta}) = \beta'\beta + \sigma^2\sum_{i=1}^{p}\lambda_i^{-1}$, where $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the eigenvalues of $X'X$ and $\hat{\sigma}^2$ is an unbiased estimator of σ^2.

One might argue that $\hat{\beta}'\hat{\beta}$ in $\hat{h}_1$ should be replaced by $\hat{\beta}'\hat{\beta} - \hat{\sigma}^2\sum_{i=1}^{p}\lambda_i^{-1}$ so as to make $\hat{h}_1$ approximately unbiased. When $X'X$ is in correlation form σ^2 must be quite small to produce good R^2 values. Accordingly, $\hat{k}_1^{-1}$ would remain essentially unchanged unless $X'X$ is highly ill-conditioned. In particular, the last matrix in Table 2 is just such a matrix. For that matrix, with $\hat{k}_2 = (\hat{\beta}'\hat{\beta} - \sigma^2\sum_{i=1}^{p}\lambda_i^{-1})/\hat{\sigma}^2$, the average value of $2\hat{k}_2^{-1}$ should be approximately $5\hat{k}_1^{-1}$ for $\sigma^2 = .0001$, approximately $8\hat{k}_1^{-1}$ for $\sigma^2 = .001$, and approximately $1.4\ \hat{k}_1^{-1}$ for $\sigma^2 = .005$. Notice that these are close to the minimum values in Table 2, especially for the second loss function. Therefore, especially if $X'X$ is highly ill-conditioned, $\hat{k}_2^{-1}$ or $(\hat{k}_2/2)^{-1}$ be considered if the numerator of $\hat{k}_2$ is sufficiently positive. Although the vast majority of our simulations have been for $p = 4$, limited work with $p = 3$ has revealed that $2\hat{k}_2^{-1}$ may be a good choice for k when $X'X$ is highly ill-conditioned and R^2 is very close to 1.0.

Many researchers have contended that the improvement of a biased estimator over least squares will usually be quite small. The figures in Table 2 indicate that this is not necessarily true. The absolute gain might be small since the data for the regressors is in correlation form, but could be quite sizable after converting back to raw form.

TABLE 1 - *Comparison of least squares, the minimax estimator* ($k=\hat{h}_1^{-1}$), *and the Hoerl, Kennard, and Baldwin estimator* ($k=4\hat{h}_1^{-1}$).

Well-conditioned X matrix ($\lambda_- = .7796$, $\lambda_+ = 1.4417$)

σ^2	Ave. R^2	Least Squares	$\hat{h}_1^{-1}$	$4\hat{h}_1^{-1}$
	.9954	.000410	.000410	.000410
.0001	.9976	.000396	.000396	.000395
	.9564	.005008	.004993	.004968
.001	.9753	.004349	.004334	.004295
	.5949	.044914	.044376	[illegible]
.01	.5014	.042076	.041366	.039948
	.2441	.463888	.434394	.401581
.1	.3201	.392342	.346304	.286141

Moderately ill-conditioned X matrix ($\lambda_- = .1250$, $\lambda_+ = 3.2766$)

σ^2	Ave. R^2	Least Squares	$\hat{h}_1^{-1}$	$4\hat{h}_1^{-1}$
	.9721	.001655	.001657	.001675
.0001	[illegible]	.001862	.001880	.001872
	[illegible]	.016137	[illegible]	.017929
.001	[illegible]	.017555	.017630	.016965
	[illegible]	.152402	[illegible]	.173409
.01	.9003	[illegible]	.145073	.111957
	.1276	1.67321	1.28927	.953034
.1	[illegible]	1.73359	1.21758	.617053

Very ill-conditioned X matrix ($\lambda_- = .0031$, $\lambda_+ = 3.6372$)

σ^2	Ave. R^2	Least Squares	$\hat{h}_1^{-1}$	$4\hat{h}_1^{-1}$
	.4797	.029231	.032594	.050302
.0001	.9990	.029267	.027737	.023895
	.1628	.376905	.428802	.498001
.001	.9902	.374973	.278238	.151009
	.0963	[illegible]	1.61059	[illegible]
.01	.9069	2.84197	1.35498	.458576
	.0480	28.8260	13.5135	5.66249
.1	.5179	28.2283	12.8263	4.91546

The loss is $(\hat{\beta}_{est}-\beta)'(\hat{\beta}_{est}-\beta)$. The X matrix is 40 × 4; λ_+ and λ_- denote the largest and smallest eigenvalues, respectively, of X'X. The first row of averages is for the smallest eigenvalue case; the second row for the largest [illegible] Averages [illegible] observations.

TABLE 2 - *Comparison of minimax-related ridge estimators using random norm-one β.*

Well-conditioned X matrix

σ^2	Avg. R^2	Least Squares	$\hat{h}_1^{-1}$	$2\hat{h}_1^{-1}$	$4\hat{h}_1^{-1}$	$6\hat{h}_1^{-1}$	$8\hat{h}_1^{-1}$	$10\hat{h}_1^{-1}$	Loss
.0001	.996	.0004218	1.0002	1.0007	1.0017	1.0031	1.0045	1.0062	A
		.0004020	1.0005	1.0007	1.0017	1.0030	1.0042	1.0060	B
.001	.964	.004425	0.9991	0.9984	0.9991	1.0020	1.0070	1.0142	A
		.004140	0.9988	0.9983	0.9986	1.0010	1.0056	1.0123	B
.005	.846	.02060	0.9961	0.9951	1.0015	1.0164	1.0456	1.0811	A
		.01939	0.9959	0.9948	1.0000	1.0160	1.0407	1.0743	B

Moderately Ill-conditioned X matrix

σ^2	Avg. R^2	Least Squares	$\hat{h}_1^{-1}$	$2\hat{h}_1^{-1}$	$4\hat{h}_1^{-1}$	$6\hat{h}_1^{-1}$	$8\hat{h}_1^{-1}$	$10\hat{h}_1^{-1}$	Loss
.0001	.993	.001831	1.0001	1.0005	1.0024	1.0055	1.0099	1.0136	A
		.0004198	1.0002	1.0007	1.0021	1.0045	1.0076	1.0117	B
.001	.937	.01842	0.9962	0.9957	1.0038	1.0233	1.0527	1.0917	A
		.004239	0.9979	0.9979	1.0038	1.0172	1.0377	1.0644	B
.005	.854	.0837	0.962	0.935	0.912	0.915	0.939	0.977	A
		.0196	0.974	0.959	0.944	0.954	0.974	1.005	B

Very Ill-conditioned X matrix

σ^2	Avg. R^2	Least Squares	$\hat{h}_1^{-1}$	$2\hat{h}_1^{-1}$	$4\hat{h}_1^{-1}$	$6\hat{h}_1^{-1}$	$8\hat{h}_1^{-1}$	$10\hat{h}_1^{-1}$	Loss
.0001	.987	.0326	0.982	0.979	1.000	1.055	1.129	1.221	A
		.000380	0.997	0.995	1.005	1.024	1.050	1.079	B
.001	.932	.339	0.796	0.630	0.590	0.555	0.543	0.546	A
		.00398	0.942	0.912	0.887	0.899	0.920	1.000	B
.005	.862	1.631	0.604	0.452	0.318	0.259	0.228	0.210	A
		.0199	0.874	0.824	0.769	0.749	0.739	0.990	B

600 "good" ($R^2 > .7$) observations were generated for each σ^2, matrix combination. The entries represent average loss for least squares and the ratio of average loss for each estimator to the average loss for least squares.

Loss A is $(\hat{\beta}_{est} - \beta)'(\hat{\beta}_{est} - \beta)$ Loss B is $(\hat{\beta}_{est} - \beta)'X'X(\hat{\beta}_{est} - \beta)$

$$= (\hat{Y} - X\beta)'(\hat{Y} - X\beta)$$

The matrices used here are the same as those used in TABLE 1.

4. SUMMARY

A minimax approach to linear regression was presented and some of the pertinent minimax results of an important Russian paper by Kuks and Olman were also discussed. Additional minimax results were presented including an invariance property of minimax linear regression estimators. In Section 3 some criticisms were made of the most common simulation procedure for ridge regression comparisons, and these were also illustrated in Table 1. Accordingly, a different procedure involving random β was used to produce Table 2.

The minimax regression estimators with ridge constants $\hat{\lambda}_1^{-1}$ and $\Sigma\hat{\lambda}_i^{-1}$ were seen to perform well in the major simulation study although $\Sigma\hat{\lambda}_i^{-2}$ may be a better choice, especially if $X'X$ is highly ill-conditioned.

5. ACKNOWLEDGMENTS

This research was supported in part by the Air Force Office of Scientific Research, USAF, under Grant No. F49620-79C-0125 [redacted]. The authors wish to thank the referees for their many helpful suggestions, especially concerning the presentation of the minimax theorems.

REFERENCES

[1] BIBBY, J. and TOUTENBURG, H. (1977). *Prediction and Improved Estimation in Linear Models.* New York: Wiley.

[2] BUNKE, O. (1975). Improved inference in linear models with additional information. *Math. Op. und Stat.*, 6, 817-829.

[3] DEMPSTER, A., SCHATZOFF, M., and WERMUTH, N. (1977). A simulation study of alternatives to ordinary least squares. *J. Amer. Statist. Assoc.*, 72, 77-90.

[4] GIBBONS, D. (1978). A simulation study of some ridge estimators. General Motors Research Laboratories, Research Publication GMR-2659. (Revised), September.

[5] GOLUB, G., HEATH, M., and WAHBA, G. (1977). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21, 215-223.

[6] HEMMERLE, W. J. and BRANTLE, T. F. (1978). Explicit and constrained generalized ridge estimation. *Technometrics*, 20, 109-120.

[7] HOERL, A. E., and KENNARD, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12, 55-67.

[8] HOERL, A. E., KENNARD, R. W., and BALDWIN, K. F. (1975). Ridge regression: some simulations. *Comm. in Statist.*, 4, 105-123.

[9] KUKS, J. and OLMAN, W. (1972). Minimax linear estimation of regression coefficients, II. *Izvestija Akademija Nauk Estonskoj SSR* 21, 66-72.

[10] Newhouse, J. P. and OMAN, S. D. (1971). An evaluation of ridge estimators. Rand Corporation Report No. R-716-PR.

[11] PEELE, L. (1979). Most-robust ridge regression estimators. Old Dominion University Department of Mathematical Sciences Technical Report TR 79-7, September.

[12] ROYDEN, H. (1968). *Real Analysis*. New York: Macmillan.

[13] RYAN, T. P. (1980). A new method of generating correlation matrices. *J. Statist. Comput. Simul.*, to appear.

[14] SMITH, G. and CAMPBELL, F. (1980). A critique of ridge regression methods. *J. Amer. Statist. Assoc.*, to appear.

[15] THEOBALD, C. M. (1974). Generalizations of mean squared error applied to ridge regression. *J. Roy. Statist. Soc.*, B, 36, 103-106.

nature. Traces of MSE estimates and of related direction cosine estimates can be used to augment the trace of ridge coefficients. It is the "see power" of these data analytic displays that glorifies enlistment in the ridge navy.

2. INVARIANCE

Nonsingular linear reparameterizations of a linear model do indeed leave the expectation vector invariant: $\mathbf{X}\beta = \mathbf{Z}\gamma$ for $\mathbf{Z} = \mathbf{XA}$ and $\gamma = \mathbf{A}^{-1}\beta$. But the amount of ridge shrinkage that is MSE optimal in the usual trace sense (Obenchain 1975) is a nonlinear function of the eigenvalues of $\mathbf{X'X}$ (or $\mathbf{Z'Z}$) and of the parameters, β (or γ) and σ^2. The unknown MSE optimal ridge fit is thus not invariant. Why, then, do Smith and Campbell feel that the fit "should be" invariant?

An estimation procedure will yield fits that are invariant in this sense if and only if the estimate $\mathbf{b}$ of β and the estimate $\mathbf{c}$ of γ are interrelated by the implicit formula $\mathbf{c} = \mathbf{A}^{-1}\mathbf{b}$ for every nonsingular $\mathbf{A}$. The ridge solutions that optimize the objective criteria of my earlier comments are not implicitly interrelated. And I feel that no "anxiety" worth avoiding is implied by this property.

It is well known that any procedure that yields implicitly interrelated estimates cannot have certain desirable properties. For example, Tukey (1975) considers the two-regressor case $z_1 = x_1 + ax_2$ and $z_2 = x_2$. The resulting implicit estimates are $c_1 = b_1$ and $c_2 = b_2 - ab_1$. This represents a change in the implicit estimate of the unchanged regressor and no change in the implicit estimate of the changed regressor. In this sense, no implicit coefficient can be interpreted as measuring the "effect" of only one regressor on the expected response.

Ridge estimates can be chosen to be less highly intercorrelated than are the implicit least squares estimates for an ill-conditioned parameterization. These ridge estimates are thus "a step in the right direction" in the sense of Tukey (1975) toward coefficients that can be interpreted as "effects" in the sense of Obenchain and Vinod (1974).

3. OBTAINING PREASSIGNED VALUES

Smith and Campbell [illegible] and 2.2) [illegible] the ingenious observation that if ridge regression is assumed to yield implicit estimates, then these coefficients can be made to correspond to any preassigned vector by choice of a diagonal matrix $\mathbf{A}$ in their equation (2.2).

If one considers generalized ridge regression with arbitrary origin, as in Smith and Campbell's equation (5.1), then one can indeed obtain any preassigned solution by choice of the ridge factors, $\alpha_1, \ldots, \alpha_p$. But, instead of $0 \leq \alpha_i \leq 1$, one might have to go to the extreme of using ridge factors greater than one or even negative. If one is monitoring sample information on the variance-bias trade-off, then one would see that MSE is not reduced by such an extreme tactic.

REFERENCES

Efron, B., and Morris, C. (1977), "Comment," *Journal of the American Statistical Association*, 72, 91–93.

Mallows, C.L. (1973), "Some Comments on C_p," *Technometrics*, 15, 661–675.

McCabe, G.P. (1978), "Evaluation of Regression Coefficients Using α-Acceptability," *Technometrics*, 20, 131–139.

Obenchain, R.L. (1975), "Ridge Analysis Following a Preliminary Test of the Shrunken Hypothesis," *Technometrics*, 17, 431–441 (with discussion by G.C. McDonald, 443–445).

——— (1977), "Classical *F*-tests and Confidence Regions for Ridge Regression," *Technometrics*, 19, 429–439.

——— (1978), "Good and Optimal Ridge Estimators," *Annals of Statistics*, 6, 1111–1121.

Obenchain, R.L., and Vinod, H.D. (1974), "Estimates of Partial Derivatives From Ridge Regression on Ill-Conditioned Data," NBER-NSF Seminar on Bayesian Inference in Econometrics, Ann Arbor, Mich.

Tukey, J.W. (1975), "Instead of Gauss-Markov Least Squares, What?," in *Applied Statistics*, ed. R.P. Gupta, Amsterdam: North-Holland Publishing Co., 351–372.

Comment

LAWRENCE C. PEELE and THOMAS P. RYAN*

The authors have contributed some interesting criticisms but their contentions seem not to form a basis for general condemnation of ridge regression. Smith and Campbell state that they advocate the use of prior information, but are uncomfortable with exact linear constraints. They also state that "when the least squares estimates are imprecise, auxiliary information is quite useful. Pseudoinformation is of dubious value."

[*] Lawrence C. Peele is Assistant Professor in the Department of Mathematical Sciences, Old Dominion University, Norfolk, VA 23508. Thomas P. Ryan is Assistant Professor in the Department of Quantitative Analysis, University of Cincinnati, Cincinnati, OH 45221. Research was supported in part by the Air Force Office of Scientific Research, Contract No. AFOSR F49620-79-C-0125.

© Journal of the American Statistical Association
March 1980, Volume 75, Number 369
Theory and Methods Section

We agree that exact linear constraints are "discomforting" (and seemingly unrealistic, as well) and, consequently, prefer constraints of the form $\beta'\beta \leq h$, or more generally $\beta'T\beta \leq h\sigma^2$, where T is a positive definite matrix and h is a positive constant. There is no boundedness assumption on $\beta'\beta$ inherent in the least squares procedure, and, in practice, one might set an a priori upper bound on reasonable values of $\beta'\beta$. The degree of superiority of a biased estimator over least squares will naturally depend on how sharp a bound on $\beta'\beta$ an experimenter can impose. Suppose that based on prior information only, one believes that $\sigma^{-2}\beta'\beta$ is approximately equal to a number k^*. Kuks and Olman (1972) showed that if $k^* \geq \sigma^{-2}\beta'\beta/2$, then the ridge regression estimator with ridge parameter $k = 1/k^*$ will outperform the least squares estimator in terms of mean squared error. Consequently, if fairly accurate prior information about $\sigma^{-2}\beta'\beta$ is available, this ridge estimator is preferable to the least squares estimator. Lacking prior information, we can still obtain, by estimating $\sigma^{-2}\beta'\beta$ from sample data, an estimator that tends to be better than least squares. It follows from this line of reasoning that estimators with this underlying motivation will not have "loose theoretical underpinnings."

Smith and Campbell note that ridge regression estimators are not invariant under model transformations of the form

$$Y = (XA)(A^{-1}\beta) + \epsilon = Z\gamma + \epsilon . \tag{1}$$

Peele and Ryan (1979) discuss minimax linear estimation based on prior information of the form $\beta'T\beta \leq \sigma^2 h$, which includes ordinary and generalized ridge regression estimation. The authors observe that the resulting estimators

$$\hat{\beta}^*_{T,h} = \left[\frac{1}{h}T + X'X\right]^{-1}X'Y$$

are invariant under model transformations of the form (1).

In his contribution to this discussion, Van Nostrand briefly mentioned specific (conflicting) results from several simulation studies. It is not surprising that the results can differ greatly from one study to the next since no "standard" simulation procedure is being used. In point of fact, it seems that those of us who have performed these studies have fallen to a number of pitfalls. In a regression simulation study, one should use appropriate values of the error variance σ^2. The use of correlation form implies that one should choose σ^2 close to zero in order to generate data with reasonable R^2 values. Unfortunately, R^2 is quite sensitive to small changes in σ^2; consequently, great care must be exercised in order to avoid producing small R^2 values. Also, the usual procedure of letting β be represented by the eigenvectors corresponding to the largest and smallest eigenvalues of $X'X$ is objectionable. In the smallest eigenvalue case, R^2 will be approximately p/n when the smallest eigenvalue is quite small and σ^2 is at least one order of magnitude larger. Thus, with $p = 4$ and $n = 100$ (a typical choice) we could hardly say that we have generated representative regression data when R^2 is approximately .04. Making σ^2 smaller in an effort to remedy the problem would only cause R^2 to be almost exactly 1.0 in the largest eigenvalue case. There is clearly a need for a better approach.

REFERENCES

Kuks, J., and Olman, W. (1972), "Minimax Linear Estimation of Regression Coefficients, II," *Izvestija Akademija Nauk Estonskoj SSR* 21, 66–72.

Peele, L., and Ryan, T. (1979), "Minimax Regression Estimators With Application to Ridge Regression," submitted to *Technometrics*.

Comment

H. D. VINOD*

Certain prior knowledge about intrinsic measurement errors can be cleverly incorporated in ridge regression by adding p fictitious observations to the data. Consider a standardized regression model $y = X\beta + u$, where $X'X$ is the correlation matrix based on n observations among p regressors and $X'y$ is a vector of correlation coefficients with the dependent variable. There are measurement errors in each regressor, so that the available data are indistinguishable from other data where one may add a number between −.5 to .499 beyond the last published digit. There may be a similar error in each mean $(\bar{x}_i)$ and standard deviation $(\mathrm{SD})_i$. The largest measurement error $(e_i^{\max})$ in each standardized regressor $(x_{ti} - \bar{x}_i)(n-1)^{-1/2}(\mathrm{SD})_i^{-1}$, where $t = 1, 2, \ldots, n$, is

* H. D. Vinod is Supervisor, Economic Studies, Economic Analysis Section, American Telephone and Telegraph Co., Piscataway, NJ 08854.

© Journal of the American Statistical Association
March 1980, Volume 75, Number 369
Theory and Methods Section

DATE

FILMED

— 8